Automatic lameness detection based on consecutive 3D-video recordings

T. Van Hertem[a,c], S. Viazzi[c], M. Steensels[a,c], E. Maltz[a], A. Antler[a], V. Alchanatis[a], A. A. Schlageter-Tello[b], C. Lokhorst[b], C. E. B. Romanini[c], C. Bahr[c], D. Berckmans[c], and I. Halachmi[a1]

[a]*Institute of Agricultural Engineering - Agricultural Research Organization (ARO) - the Volcani Center, PO Box 6, Bet-Dagan IL-50250, Israel*

[b]*WageningenUR Livestock Research, P.O. Box 65, NL-8200AB Lelystad, Netherlands*

[c]*Division M3-BIORES: Measure, Model & Manage Bioresponses, KU Leuven, Kasteelpark Arenberg 30 - bus 2456, BE-3001 Leuven, Belgium*

Corresponding author:

Ilan Halachmi

Institute of Agricultural Engineering – Agricultural Research Organization (ARO) – the Volcani Center,

P.O.Box 6, Bet Dagan IL-50250, Israel

Email: halachmi@volcani.agri.gov.il

fax.: +972 4 9539 566

tel.: +972 50 6220 112

[1] Corresponding author: halachmi@volcani.agri.gov.il

ABSTRACT

Manual locomotion scoring for lameness detection is a time-consuming and subjective procedure. Therefore, the objective of this study is to optimise the classification output of a computer vision based algorithm for automated lameness scoring. Cow gait recordings were made during four consecutive night-time milking sessions on an Israeli dairy farm, using a 3D-camera. A live on-the-spot assessed 5-point locomotion score was the reference for the automatic lameness score evaluation. A dataset of 186 cows with four automatic lameness scores and four live locomotion score repetitions was used for testing three different classification methods.

The analysis of the automatic scores as independent observations led to a correct classification rate of 53.0% on a 5-point level scale. A multinomial logistic regression model based on four individual consecutive measures obtained a correct classification rate of 60.2%. When allowing a 1 unit error on the 5-point level scale, a correct classification rate of 90.9% was obtained. Strict binary classification to Lame vs. Not-Lame categories reached 81.2% correct classification rate.

The use of cow individual consecutive measurements improved the correct classification rate of an automatic lameness detection system.

Key words. dairy cow, lameness, computer vision, classification, consecutive measurements

1. Introduction

Lameness is a major welfare issue in modern intensive dairy farming (Lievaart and Noordhuizen, 2011; Bruijnis, Beerda, Hogeveen & Stassen, 2012). Liveweight (Alawneh, Stevenson, Williamson, Lopez-Villalobos et al., 2012), milk fat (van Straten, Siani & Bar, 2011), milk yield (Green, Hedges, Schukken, Blowey et al., 2002; Archer, Green & Huxley, 2010; Reader, Green, Kaler, Mason et al., 2011) and behavioural indicators such as activity and lying time (Blackie, Amory, Bleach & Scaife, 2011; Pavlenko, Bergsten, Ekesbo, Kaart et al., 2011; Reader et al., 2011) were all affected when a lameness incident occurred. Prevalence rates are influenced by housing (Potterton, Green, Harris, Millar et al., 2011), farm management practices (Chapinal, Barrientos, von Keyserlingk, Galo et al., 2013), feed (Amory, Kloosterman, Barker, Wright et al., 2006) and breed (Barker, Leach, Whay, Bell et al., 2010). Averaged reported lameness prevalence rates range from 20% to 25% in USA, and 33% to 37% in Europe (Schlageter Tello, Bokkers, Groot Koerkamp, Van Hertem et al., 2013a). Herd locomotion scoring is a common method to obtain a lameness prevalence rate (Flower & Weary, 2009). However, this procedure is time-consuming and subjective. It is therefore hardly done in practice, and when done for large herd sizes, it is often done on a subsample of the entire herd (Thomsen, 2009; Main, Barker, Leach, Bell et al., 2010).

In the scientific community, different approaches have been developed to automate locomotion scoring and lameness detection. Studies using force plates (Pastell, Hautala, Poikalainen, Praks et al., 2008; Liu, Dyer, Neerchal, Tasch et al., 2011; Ghotoorlar, Ghamsari, Nowrouzian & Ghidary, 2012), pressure sensitive walkways (Maertens, Vangeyte, Baert, Jantuan et al., 2011) and accelerometers (Pastell, Tiusanen, Hakojarvi & Hanninen, 2009; Chapinal, De Passille,

Pastell, Hanninen et al., 2011) have measured the asymmetry of the gait when walking. These approaches used the diversity in weight bearing on each leg for lameness detection.

The use of (mostly behavioural) sensors is spreading in modern dairy farming and there are studies that use existing sensor data on the farm such as lying time (Ito, von Keyserlingk, LeBlanc & Weary, 2010), feeding behaviour and neck activity (Kramer, Cavero, Stamer & Krieter, 2009), and ruminating time, neck activity and milk yield (Van Hertem, Maltz, Antler, Schlageter Tello et al., 2013) for lameness detection.

Other studies have used two-dimensional (2D) computer vision to analyse gait. These studies focused on the measurement of different gait and posture variables such as back arch curvature (Poursaberi, Bahr, Pluk, Van Nuffel et al., 2010; Viazzi, Bahr, Schlageter Tello, Van Hertem et al., 2014), step overlap (Pluk, Bahr, Leroy, Poursaberi et al., 2010), hoof release angles (Pluk, Bahr, Poursaberi, Maertens et al., 2012), the body movement pattern (Poursaberi, Bahr, Pluk, Veermae et al., 2011). The challenge to practical application of this method is to identify accurately the location in the image of anatomical body parts such as hooves, limb joints, withers and back contour lines. Until now, this has been performed using manually labelled markers attached to the limbs of the cows (Aoki, Kamo, Kawamoto, Zhang et al., 2006; Song, Leroy, Vranken, Maertens et al., 2008; Blackie, Bleach, Amory & Scaife, 2011). The manual labelling step inhibits full automation. Video pre-processing provides an alternative for locating the anatomical body parts in the video. During pre-processing, videos are transformed to sequences of binary images in which anatomical parts of cows can be clearly segmented from the background. Van Hertem, Alchanatis, Antler, Maltz et al. (2013) showed however that image segmentation in 2D RGB-images was problematic in real farm conditions due to dynamic background restrictions in side view perspective. To overcome these restrictions, a three-

dimensional (3D) camera in top view perspective was suggested to obtain the same body movement pattern as with the side view camera. Viazzi, Bahr, Van Hertem, Schlageter Tello et al. (2014) developed an automated lameness scoring algorithm based on 3D imaging of the cow's gait. The algorithm calculated the back posture measurement, which was related to the back arch curvature, one of the key indicators of cow lameness (Sprecher, Hostetler & Kaneene, 1997). The algorithm was validated on a small dataset of 92 cows and obtained an accuracy of 90% (Viazzi, Bahr, Van Hertem, et al., 2014).

In order to cope with false alarms – for instance when a cow trips or slips - an optimisation of the algorithm is necessary. A decision made over several consecutive measurements should reduce false alarms and optimise classification performance. Therefore, the aim of this study was to improve the algorithm developed by Viazzi, Bahr, Van Hertem, et al., (2014) and to optimise the classification performance by taking into account consecutive measurements.

2. Materials and methods

2.1.Nomenclature

- BPM: Back Posture Measurement;
- LS: Locomotion Score;
- AS: Algorithm Score;
- 5-point classification: the classification when model and reference are subdivided in five classes;
- Binary classification: the classification when model and reference are transformed from the 5-point classification to a Lame vs. Not Lame scale;
- 'Strict': the exact agreement between model and reference;

- ‘Tolerant’: the acceptable one single unit difference between model and reference;
- CCR: correct classification rate or model accuracy;
- MAE: mean absolute error;
- RMSD: root mean squared deviation;
- CC: contingency coefficient;
- Sensitivity: the ability to correctly classify Lame cows;
- Specificity: the ability to correctly classify Not Lame cows;
- MCR: the tolerant misclassification rate or the misclassifications with a difference between model and reference bigger than one unit.

2.2. Animals and housing

Data were obtained by the Agricultural Research Organization (ARO) in a commercial dairy farm in kibbutz Yifat, Israel. The milking herd consisted of 1100 Israeli-Holstein dairy cows. The herd contained 12 production groups according to parity, lactation stage, reproduction status and health status (group size: 96 ± 12 cows). All cows were housed in separate no-stall fully roofed open cowsheds with dried manure bedding material. Lameness prevalence ranged between 15% and 25% throughout the year.

All cows were milked three times a day (start at 0300 h, 1100 h and 1900 h) in a 2x32 side-by-side milking parlour. The production groups were brought one-by-one to the waiting parlour to avoid group mixing. All cows were fed a total mixed ration according to NRC recommendations (NRC, 2001). The last six groups of the milking herd (approximately 500 cows) were followed. These groups consisted of multiparous cows and lameness prevalence was highest in these

groups. The same procedure was done by Viazzi, Bahr, Van Hertem et al. (2014) on the same farm.

2.3. 3D camera setup

An after-milking sorting gate was the only place on the farm that all cows passed. The 3D-camera was located after the sorting gate. A wide lane allowed heavy cow traffic to the milking parlour. In order to make the cows walk in the camera field of view, a mobile narrow corridor (maximum width = 2.10 m; minimum width = 1.00 m) was built directly after the sorting gate (Figure 1). A static corridor would interfere with the bi-weekly manure scraping in the lanes. The sorting gate and the 90^{o} turn in the corridor provided the necessary time delay between successive cows. In the corridor, the cows walked in a straight line and behind each other. The corridor was built to obtain smooth cow traffic and to avoid queuing as much as possible in real farm conditions.

(Insert Figure 1 here)

Cow gait was recorded with a Microsoft Kinect Xbox 3D-camera (Kinect™, Microsoft corp., Washington, USA). The camera was positioned in top down perspective, 3.20 m above ground level (Figure 1), in order to have the full cow body (head-tail length range: 2.44 m – 2.80 m) visible in the camera field of view at 30 fps. A photocell (HRTL 96B™, Leuze electronic GmBH, Owen, Germany) was used to trigger the video recording. This photocell was located 0.5 m before the beginning of the camera field of view, and was linked to a programmable logic controller (NI USB-6501, National Instruments, Austin Texas, USA). The controller was set to record four seconds in order to have only one cow per video. The camera was connected to an

operating computer through a USB-port. After each data collection session, the construction was packed away along the walking lane path, where it did not interfere with the farm routine. The recorded videos contained a depth recording (for 3D-reconstruction) and a RGB-recording and were saved as .oni-files to a 1TB hard disk (Western Digital, Irvine California, USA). The OpenNI 1.0 Software Development Kit framework (www.openni.org, last accessed at 24 March 2012) was used to make recordings with the Kinect camera.

Due to the sunlight sensitivity of the cameras, data were collected during four consecutive night milking sessions. External artificial light sources were installed around the video corridor, but not pointing directly to the sensor, to increase cow visibility for locomotion scoring and visual identification.

2.4.Locomotion score reference

During each data collection period, the locomotion of cows passing the corridor was manually scored on-the-spot by the same one trained observer (so called *'locomotion score'*, *'LS'* or *'reference'*). The locomotion scoring was based on the discrete 5-point numerical score of Sprecher, Hostetler and Kaneene (1997) [1=healthy; 5= severely lame].

The intra-observer repeatability of the trained observer for the 5-point score was quantified and achieved a kappa coefficient (based on Cohen (1960)) of 0.53 ± 0.02, and a weighed kappa coefficient of 0.69 ± 0.03 in the four consecutive locomotion scoring sessions. These intra-observer repeatability values indicate substantial repeatability (Landis & Koch, 1977), and proof of the level of training.

2.5. 3D video analysis

All recorded videos were analysed automatically with a software program. The software included the algorithm for the extraction of the four image features described by Viazzi, Bahr,

Van Hertem et al. (2014) from the depth images. On top of that, the four parameters were combined in one output variable "Back Posture Measurement" (BPM) by a weighing function similar to the function described in Viazzi, Bahr, Van Hertem et al. (2014) for 2D-images. The software was compiled with a MATLAB Runtime Compiler (Matlab® R2011b, The MathWorks©, Inc, Natick, MA, United States) and executed on the recording computer after video recording. The input of this software was the Kinect output video in .oni-format. The software analysed all depth-frames in which a full cow body shape could be segmented. On average 4.9 ± 2.7 frames per video were analysed, depending on the walking speed of the cow. The output was a MATLAB-file including all usable depth-frames for analysis, together with the associated RGB-frames, and the generated BPM-value related to the cow's back posture in the video. The BPM-output was calculated as the median value of all frames in the video that were stored in the MATLAB-file. The BPM-output in this herd ranged from 0.13 to 0.33..

The cows in the video were manually identified based on the recorded RGB-video frames. The video timestamp was used to double check cow identification with the automatically generated cow list in the sorting gate when the numbers on the back were hard to read.

At the end of the analysis, a report was generated containing cow number, the number of usable frames in the video related to the cow number, and the BPM-value.

2.6.Data selection

In four consecutive night milking sessions, 1327 complete cow-observations were done on 511 individual cows. A complete observation consisted of a live locomotion score by the observer and a successfully recorded video. A subset of 186 individual cows that were identified in all four consecutive sessions (4x186 = 744 cow-observations) was selected for further analysis. The three selection criteria were that (i) four consecutive locomotion scores were available; (ii) four

consecutive videos were available; (iii) the four consecutive locomotion scores did not vary more than one numerical unit to reduce human errors in the reference. The first two criteria reduced the dataset to 195 individual cows and 780 cow-observations. The last criteria was not met by 9 individual cows, and therefore 36 more cow observations were omitted. This implied that 583 cow-observations (44%) were omitted for further analysis.

2.7. Classification procedures

2.7.1. **Independent cow observations.**

The AS was compared to LS. In order to put the BPM-score on the same scale as the LS [classes 1-5], a rescaling of the BPM-values was done with Equation 1, with *min(BPM)* the minimum value of BPM, and *max(BPM)* the maximum value of BPM.

$$AS = 0.5 + 5*((BPM - min(BPM)) / (max(BPM)-min(BPM))) \quad \text{(Equation 1)}$$

The AS-values were transformed to their nearest integer values.

Another approach to put the BPM-score on the same scale as the LS was using four non-equidistant cut-off thresholds. For each combination of cut-off thresholds in the range of [*min(BPM), max(BPM)*], the CCR and MCR were calculated. The four thresholds that maximised the 5-point CCR were selected as the maximising CCR-thresholds. The thresholds that minimised the 5-point MCR were selected as the minimising MCR-thresholds.

2.7.2. **Models with individual consecutive measurements.**

Instead of analysing all measurements as independent observations, the 744 cow-observations were considered as four consecutive measures of the same 186 individual cows. Three different classification models were developed and compared to each other, by using the MATLAB

R2011b Statistics Toolbox. These models take into account multiple consecutive readings before making a final classification.

Classification models. For model calibration, 2/3 of the data were used, and the remaining 1/3 of the data were used for model validation. The model calibration and validation dataset had an equal proportion and distribution according to the reference data. The rounded average rounds the decimal number to the nearest integer value. If the decimal of the average value of four consecutive measurements was equal to 0.5, the rounded average was rounded downward if the last measurement was equal to the lowest value in the range, and upwards if equal to the highest value in the range.

- An *ordinal multinomial logistic regression model* was used because the rounded average of four consecutive locomotion scores, used as reference, was interpreted as an ordinal outcome variable according to lameness severity. The model allows multiple discrete outcomes, in order to predict the probabilities of the different possible outcomes of a categorical outcome variable, given a set of four consecutive BPM measurements as input variables (Hosmer & Lemeshow, 2000).
- When a *nominal multinomial logistic regression model* was used, the rounded average of four consecutive locomotion scores, used as reference, was interpreted as a nominal outcome variable with no relationship between the different categorical classes. The model allows multiple discrete outcomes, in order to predict the probabilities of the different possible outcomes of a categorical outcome variable, given a set of four consecutive BPM-scores as input variables (Hosmer & Lemeshow, 2000).
- A *linear regression model* was used when the reference, the average of four consecutive locomotion scores, was considered as a continuous variable. The model makes a

weighted sum of the four consecutive BPM input values, in such a way that it would fit the output variable. The weights are determined by applying the least squares method to the calibration dataset (Neter, Kutner, Nachtsheim & Wasserman, 1996).

Improving model robustness. In order to obtain more robust model outcomes, a cross-validation and a bootstrap aggregating procedure were used.

- All models were validated using a 5-fold (k = 5) cross-validation procedure on the 186 cow repetitions. The number of folds in the cross-validation process had to be limited to the level of five otherwise not all folds would have cows in category five. Each fold contained an equal proportion and distribution according to the reference data.
- A bootstrap aggregating (bagging) procedure is applied to the 5-fold cross-validated classification model. Bagging is a model-averaging approach to improve the machine learning of statistical classification models regarding stability and classification accuracy, variance reduction and avoiding over-fitting (Breiman, 1996). Starting from a standard training dataset, bagging generates m (m = 200) new training sets, each of size s (s = 150) which is smaller than the size of the initial standard training dataset n (n = 186), by sampling observations from the initial dataset uniformly and with replacement. By sampling with replacement, some samples can be duplicated in each m. The m models are fitted using the s samples and combined by voting. For discrete outcomes, a voting procedure counts the number of votes each discrete outcome received after the m models were developed.

2.7.3. Confusion matrix

A confusion matrix was used to evaluate the classification model output against the LS reference.

5-point classification. In the 5-point confusion matrix, both the reference and the model output are tabulated in 5 levels. The CCR is defined as the sum of the elements on the main diagonal in the confusion matrix. Besides CCR, the 5-point classification performance is also expressed by the mean absolute error (MAE), the root mean squared deviation (RMSD) and the contingency coefficient (CC) of the confusion matrix.

Strict binary classification. The 5-point locomotion score was transformed to a binary score (Lame vs. Not-Lame). Cows that were scored as LS = 1 or LS = 2, were considered to be Not-Lame, and cows that were scored as LS = 3, LS = 4 or LS = 5 were considered to be Lame.

3. Results

3.1. Classification of independent cow-observations

- Applying thresholds that maximise 5-point CCR (T1 = 0.16, T2 = 0.21, T3 = 0.25 and T4 = 0.30) resulted in a correct classification rate of 53.0% (Table 1a).
- Applying thresholds that minimise 5-point MCR (T1 = 0.14, T2 = 0.24, T3 = 0.26 and T4 = 0.30) resulted in a correct classification rate of 42.6% (Table 1b).
- Applying equidistant thresholds after rescaling (T1 = 0.17, T2 = 0.21, T3 = 0.25 and T4 = 0.29) resulted in a correct classification rate of 52.6% (Table 1c).

(Insert Table 1 here)

3.2. Bootstrap voting classification of the nominal multinomial logistic regression model with individual consecutive measurements

Strict classification. The 5-point confusion matrix is presented in Table 2. Bootstrap voting on the 5-fold cross validated ordinal multinomial logistic regression model resulted in a CCR of 60.2% (Table 4a).

(Insert Table 2 here)

Strict binary classification results are presented in Table 3, and reached a CCR of 81.2% (Table 4f). Sensitivity of strict binary classification was 47.1%, and specificity was 94.1% (Table 4f).

(Insert Table 3 here)

Tolerant classification. When allowing a 1 unit error, the 5-point CCR was 90.9% (Table 4b). The MAE is 0.500, the RMSD is equal to 0.852 and the CC is equal to 0.791.

3.3. Classification performance of three regression models with consecutive measurements

5-point classification. The tolerant CCR was comparable for all three models (90.9%, 91.4% and 91.9%; Table 4b). The linear regression model had the lowest strict CCR (56.5%; Table 4a) compared to both multinomial logistic regression models (60.2%; Table 4a).

Binary classification. Regarding the strict binary classification of lameness, the ordinal multinomial logistic regression model had a higher CCR (81.2%; Table 4f) than the nominal

304 multinomial logistic regression model (80.7%) and the linear regression model (80.7%). The best
305 sensitivity value was obtained with the linear regression model (54.9%), which was higher than
306 the sensitivity of the nominal and ordinal multinomial logistic regression models (47.1%). Model
307 specificity was higher for the ordinal multinomial logistic regression model (94.1%) than for the
308 nominal multinomial logistic regression model (93.3%) and the linear regression model (90.4%).
309
310 (Insert Table 4 here)

311 *4. Discussion*

312 The results suggest that an accurate lameness detection can be made by applying a 3D-camera
313 and a multinomial logistic regression of four consecutive measurements. Independent analysis of
314 the 3D-algorithm output was not good enough (CCR = 53.0% on a 5-point level).
315 Taking into account four measurements by applying a classification model such as the ordinal
316 multinomial logistic regression, the correct classification rate was improved to 60.2%. Three
317 different classification models were tested in this study, and they differed slightly in handling the
318 reference. The four consecutive measurements were considered as dependent repetitions of the
319 same individual cow, while in the first approach, all measurements were analysed independently.
320 Consecutive measurement analysis requires an established database while independent analysis
321 can deliver answers immediately on-the-spot without the need to establish a database.
322 Making use of consecutive measurements increased the certainty of the model in order to avoid
323 presenting false alarms to the farmers. De Mol, Bleumer, van der Werf and Van Reenen (2012)
324 used a similar approach based on seven consecutive days. In this study, neither the optimal
325 classification model settings nor the optimal number of consecutive measurements were tested
326 for obtaining the best classification results. In further research, the independent analysis should

be further developed, perhaps after improving the BPM value by applying other filters and image processing techniques.

Improvement of CCR from 60.2% to 81.2% was achieved when transforming the 5-point scale to a strict binary scale (Lame vs. Not-Lame), a method previously applied by Winckler and Willen (2001), Channon, Walker, Pfau, Sheldon et al. (2009) and Main et al. (2010). A binary score is simple and easy to understand and it gives an agronomic value to the algorithm output. For practical use, the farmer needs to be informed about which cows are lame and need treatment, and which are not lame. On the other hand, a binary classification hides some useful information. A commonly used cut-off threshold to differentiate between clinical and subclinical lame cows is between 2 and 3 (Winckler & Willen, 2001). Cows that were scored as 2 or 3 by the reference or the model had a larger impact on the strict classification, whereas for the tolerant classification these values were still acceptable.

In the presented analysis, cut-off thresholds were determined at a group level. In further research when more consecutive measurements can be available, cow-specific individual cut-off thresholds should be calculated. Viazzi, Bahr, Schlageter-Tello et al. (2013) have shown that using an individual threshold on a BPM time series can improve the accuracy of the model by 10%.

In this study, a 5-point live locomotion scoring (Sprecher, Hostetler & Kaneene, 1997) was performed on-the-spot, and served as the reference. This ‘gold’ standard is known to be subjective and inter- and intra-observer repeatability is low (Flower & Weary, 2009; Schlageter Tello, Bokkers, Groot Koerkamp, Van Hertem et al., 2013b). Locomotion scoring is however a commonly used method because it provides an immediate, on-site assessment and it does not require technical equipment (Flower & Weary, 2009). As a first tentative approach to achieve

higher reliability in the reference and reduce the subjectivity effect of the scorer, only cows that had four consecutive live scorings within a one unit score-range were selected (n = 186 cows) before the analysis. A second tentative suggestion was allowing a 1 unit error in the 5-point scale, and this led to a tolerant CCR of 90.9% (n = 169 cows). A MAE of 0.50 shows that the 5-point classification is acceptable.

The accuracy of the strict binary classification, CCR = 81.2%, is lower than the accuracy obtained by Viazzi, Bahr, Van Hertem, et al. (2014), who achieved a CCR of 90%. Viazzi, Bahr, Van Hertem, et al. (2014) gathered their data in one single evening milking session, whereas the data in this analysis were gathered on four consecutive evening milking sessions. This may suggest that the daily reinstallation of the camera might influence the recording.

The comparison between the three classification models revealed that the linear regression model (54.9% vs. 47.1% for the multinomial logistic regression models) obtained higher sensitivity values on a binary scale than the multinomial logistic regression models. For the 5-point scale however, the multinomial logistic regression model (60.2%) performed better than the linear regression model (56.5%). Depending on the desired outcome (5-point scale or binary scale), one can choose the best classification model.

It is important to compare the results in our study with other studies. Care should be taken when comparing tolerant results to strict outcomes since they are not the same. Tolerant analysis allows a one unit error in the outcome, whereas the strict analysis only allows the exact agreement between the outcome and the reference. The CCR in our study (CCR = 90.9%) was higher in comparison to the GAITWISE system (Maertens et al., 2011) (CCR = 133/159 = 83.7%) which was based on kinematic variables. GAITWISE reached a model sensitivity of 76-90%, and a model specificity of 86-100% for a three level gait score. In the setup, they needed a

large separation between the animals in order not to have two cows on the pressure sensitive walkway at the same time. In our study, a separation between the animals was also necessary, but a small separation was sufficient to automatically differentiate consecutive cows during image segmentation. GAITWISE's lowest sensitivity values were obtained for the middle class scored cows (76%), indicating the biggest difficulty is detecting the mildly lame animals. These results compare to the results in this study, where the cows scored as LS = 2 and LS = 3 have a large impact on binary model classification accuracy.

Poursaberi et al. (2010) obtained an accuracy of 96% on a dataset of 184 cows. Their algorithm analysed the back curvature with the inverse radius variable on 2D side view images. Their side view video recordings however were made in controlled experimental conditions on an experimental dairy farm - not on a commercial dairy farm.

In this semi-automatic setup, the video recording and analysis were done fully automatically. The identification of the cows in the videos was however done manually, after the videos were recorded. A time delay between the cows is only necessary for the image processing in the analysis phase. The time to analyse each individual video is rather small (< 10 seconds). In a fully automated setup (recording + identification + analysis) that is also integrated into the farm management software, it would be possible to operate a separation fence to sort the lame cows from the non-lame cows. It is however advised to have some distance (at this moment the ideal distance is unknown) between the recording spot and the separation fence. This allows more time for recorded video analysis, transforming the system outcome to a sorting gate input signal and hence smoother cow traffic . The real implementation in farm conditions and what is the best way to use the system should be further investigated.

The recordings in this study were made during night time milking sessions due to the sunlight sensitivity of the camera. The corridor was built in an unroofed part of the farm. If however night milking sessions are not available, recordings could also be made in shaded areas where the direct or diffuse sunlight is low, because it is the infrared spectrum of the sunlight that affects camera performance. The system is therefore also applicable in other dairy husbandry systems.

The automatic scoring in our study only focused on the arching of the cow's back. Detecting only the back arching as a method to detect lameness has been described earlier by Poursaberi et al. (2010), Poursaberi et al. (2011) and Viazzi, Bahr, Schlageter Tello, et al. (2014). However, when performing a locomotion scoring, the back arching is only one of the indicators (Flower & Weary, 2009; Schlageter Tello, Lokhorst, Van Hertem, Halachmi et al., 2011). In further research, it is advised that other parameters such as gait asymmetry and head bob should also be included in the 3D-video analysis.

Van Hertem et al. (2013b) developed an automatic lameness detection model based on behavioural and performance variables that reached 85% model sensitivity and 89% model specificity. Future research is needed to reveal whether the combination of both approaches (computer vision, and behaviour and performance sensing) increases the classification accuracy of the combined model, and which variables should be included in the combined model.

5. Conclusions

A 3D-video based algorithm for lameness detection was validated in real farm conditions and compared with live locomotion scoring.

- Independent cow-observation analysis resulted in a correct classification rate of 53.0%.
- When four individual consecutive measurements were taken into account in a multinomial logistic regression model, a correct classification rate of 60.2% was reached.

- After transforming the algorithm output to a strict binary scale in order to give it a biological meaning, a correct classification rate of 81.2% was obtained.
- When allowing a one unit error as a tentative approach to reduce observer-related errors, a tolerant correct classification rate of 90.9% was obtained.

The above-mentioned results show that the use of consecutive measurements improve the classification output of a computer vision system. This conclusion should be considered before implementing an automatic lameness detection system on more farms.

6. Acknowledgements

The authors thank all farm personnel for their help on the farm. The authors thank Uzi Birk and Daniel Rozen from DeLaval for their help with the product development. The help from Doron Bar, Rony Meir and staff from SCR is also very much appreciated. The authors thank V. Ostrovsky for his help with the setup.

This study is part of the Marie Curie Initial Training Network BioBusiness (FP7-PEOPLE-ITN-2008). This study is contribution number 459-4398-951, funded by the Agricultural Research Organization (ARO), P.O.Box 6, Bet Dagan, Israel.

7. References

Alawneh, J. I., M. A. Stevenson, N. B. Williamson, N. Lopez-Villalobos & T. Otley. 2012. The effect of clinical lameness on liveweight in a seasonally calving, pasture-fed dairy herd. Journal of Dairy Science 95(2):663-669.

Amory, J. R., P. Kloosterman, Z. E. Barker, J. L. Wright, R. W. Blowey, et al. 2006. Risk factors for reduced locomotion in dairy cattle on nineteen farms in the Netherlands. Journal of Dairy Science 89(5):1509-1515.

Aoki, Y., M. Kamo, H. Kawamoto, J. Zhang & A. Yamada. 2006. Changes in walking parameters of milking cows after hoof trimming. Animal Science Journal 77(1):103-109.

Archer, S. C., M. J. Green & J. N. Huxley. 2010. Association between milk yield and serial locomotion score assessments in UK dairy cows. Journal of Dairy Science 93(9):4045-4053.

444 Barker, Z. E., K. A. Leach, H. R. Whay, N. J. Bell & D. C. J. Main. 2010. Assessment of
445 lameness prevalence and associated risk factors in dairy herds in England and Wales. Journal
446 of Dairy Science 93(3):932-941.
447 Blackie, N., J. R. Amory, E. Bleach & J. Scaife. 2011. The effect of lameness on lying behaviour
448 of zero grazed Holstein dairy cattle. Applied Animal Behaviour Science 134(3):85-91.
449 Blackie, N., E. Bleach, J. R. Amory & J. Scaife. 2011. Impact of lameness on gait characteristics
450 and lying behaviour of zero grazed dairy cattle in early lactation. Applied Animal Behaviour
451 Science 129(2-4):67-73.
452 Breiman, L. 1996. Bagging predictors. Machine learning 24(2):123-140.
453 Bruijnis, M. R. N., B. Beerda, H. Hogeveen & E. N. Stassen. 2012. Assessing the welfare impact
454 of foot disorders in dairy cattle by a modeling approach. Animal 6(6):962-970.
455 Channon, A. J., A. M. Walker, T. Pfau, I. M. Sheldon & A. M. Wilson. 2009. Variability of
456 Manson and Leaver locomotion scores assigned to dairy cows by different observers.
457 Veterinary Record 164(13):388-392.
458 Chapinal, N., A. K. Barrientos, M. A. G. von Keyserlingk, E. Galo & D. M. Weary. 2013. Herd-
459 level factors for lameness in freestall farms in the northeastern United States and California.
460 Journal of Dairy Science 96(1):318-328.
461 Chapinal, N., A. M. De Passille, M. Pastell, L. Hanninen, L. Munksgaard, et al. 2011.
462 Measurement of acceleration while walking as an automated method for gait assessment in
463 dairy cattle. Journal of Dairy Science 94(6):2895-2901.
464 Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological
465 Measurement 20(1):37-46.
466 de Mol, R. M., A. G. Bleumer, J. T. N. van der Werf & C. G. Van Reenen. 2012. Automated
467 detection of lameness in dairy cows based on day-to-day variation in behaviour. Page 396 in
468 63rd Annual Meeting of the European Federation of Animal Science (EAAP). Vol. Book of
469 Abstracts. Wageningen Academic Publishers, Bratislava, Slovakia.
470 Flower, F. C. & D. M. Weary. 2009. Gait assessment in dairy cattle. Animal 3(1):87-95.
471 Ghotoorlar, S. M., S. M. Ghamsari, I. Nowrouzian & S. S. Ghidary. 2012. Lameness scoring
472 system for dairy cows using force plates and artificial intelligence. Veterinary Record
473 170(5):126-153.
474 Green, L. E., V. J. Hedges, Y. H. Schukken, R. W. Blowey & A. J. Packington. 2002. The
475 impact of clinical lameness on the milk yield of dairy cows. Journal of Dairy Science
476 85(9):2250-2256.
477 Hosmer, D. W. & S. Lemeshow. 2000. Applied Logistic Regression. John Wiley & Sons, New
478 York, NY, USA.
479 Ito, K., M. A. G. von Keyserlingk, S. J. LeBlanc & D. M. Weary. 2010. Lying behavior as an
480 indicator of lameness in dairy cows. Journal of Dairy Science 93(8):3553-3560.
481 Kramer, E., D. Cavero, E. Stamer & J. Krieter. 2009. Mastitis and lameness detection in dairy
482 cows by application of fuzzy logic. Livestock Science 125(1):92-96.
483 Landis, J. R. & G. G. Koch. 1977. The measurement of observer agreement for categorical data.
484 Biometrics 33(1):159-174.
485 Lievaart, J. J. & J. P. T. M. Noordhuizen. 2011. Ranking experts' preferences regarding measures
486 and methods of assessment of welfare in dairy herds using Adaptive Conjoint Analysis.
487 Journal of Dairy Science 94(7):3420-3427.

Liu, J. B., R. M. Dyer, N. K. Neerchal, U. Tasch & P. G. Rajkondawar. 2011. Diversity in the magnitude of hind limb unloading occurs with similar forms of lameness in dairy cows. Journal of Dairy Research 78(2):168-177.

Maertens, W., J. Vangeyte, J. Baert, A. Jantuan, K. C. Mertens, et al. 2011. Development of a real time cow gait tracking and analysing tool to assess lameness using a pressure sensitive walkway: The GAITWISE system. Biosystems Engineering 110:29-39.

Main, D. C. J., Z. E. Barker, K. A. Leach, N. J. Bell, H. R. Whay, et al. 2010. Sampling strategies for monitoring lameness in dairy cattle. Journal of Dairy Science 93(5):1970-1978.

Neter, J., M. H. Kutner, C. J. Nachtsheim & W. Wasserman. 1996. Applied linear regression models. Third edition ed. The McGraw-Hill Companies, Inc.

NRC. 2001. Nutrient Requirements of Dairy Cattle. Vol. 1. 7 ed. National Academies Press, Washington D.C.

Pastell, M., M. Hautala, V. Poikalainen, J. Praks, I. Veermae, et al. 2008. Automatic observation of cow leg health using load sensors. Computers and Electronics in Agriculture 62(1):48-53.

Pastell, M., J. Tiusanen, M. Hakojarvi & L. Hanninen. 2009. A wireless accelerometer system with wavelet analysis for assessing lameness in cattle. Biosystems Engineering 104(4):545-551.

Pavlenko, A., C. Bergsten, I. Ekesbo, T. Kaart, A. Aland, et al. 2011. Influence of digital dermatitis and sole ulcer on dairy cow behaviour and milk production. Animal 5(8):1259-1269.

Pluk, A., C. Bahr, T. Leroy, A. Poursaberi, X. Song, et al. 2010. Evaluation of step overlap as an automatic measure in dairy cow locomotion. Transactions of the Asabe 53(4):1305-1312.

Pluk, A., C. Bahr, A. Poursaberi, W. Maertens, A. van Nuffel, et al. 2012. Automatic measurement of touch and release angles of the fetlock joint for lameness detection in dairy cattle using vision techniques. Journal of Dairy Science 95(4):1738-1748.

Potterton, S. L., M. J. Green, J. Harris, K. M. Millar, H. R. Whay, et al. 2011. Risk factors associated with hair loss, ulceration, and swelling at the hock in freestall-housed UK dairy herds. Journal of Dairy Science 94(6):2952-2963.

Poursaberi, A., C. Bahr, A. Pluk, A. Van Nuffel & D. Berckmans. 2010. Real-time automatic lameness detection based on back posture extraction in dairy cattle: Shape analysis of cow with image processing techniques. Computers and Electronics in Agriculture 74(1):110-119.

Poursaberi, A., C. Bahr, A. Pluk, I. Veermae, E. Kokin, et al. 2011. Online Lameness Detection in Dairy Cattle Using Body Movement Pattern. in 11th International Conference on Intelligent Systems Design and Applications (ISDA 2011).

Reader, J. D., M. J. Green, J. Kaler, S. A. Mason & L. E. Green. 2011. Effect of mobility score on milk yield and activity in dairy cattle. Journal of Dairy Science 94(10):5045-5052.

Schlageter Tello, A., E. A. M. Bokkers, P. W. G. Groot Koerkamp, T. Van Hertem, S. Viazzi, et al. 2013a. Locomotion scoring as a golden standard for automated lameness detection: a review. Preventive Veterinary Medicine *(submitted)*.

Schlageter Tello, A., E. A. M. Bokkers, P. W. G. Groot Koerkamp, T. Van Hertem, S. Viazzi, et al. 2013b. Within and between observer repeatability of live and video locomotion assessments in dairy cows. Animal Welfare *(submitted)*.

Schlageter Tello, A., C. Lokhorst, T. Van Hertem, I. Halachmi, E. Maltz, et al. 2011. Selection of a golden standard for visual-based automatic lameness detection for dairy cows. in Animal hygiene and sustainable livestock production. Proceedings of the XVth International

Congress of the International Society for Animal Hygiene. Vol. 1. J. Kofer & H. Schobesberger, ed. Tribun EU, Vienna, Austria.
Song, X. Y., T. Leroy, E. Vranken, W. Maertens, B. Sonck, et al. 2008. Automatic detection of lameness in dairy cattle - Vision-based trackway analysis in cow's locomotion. Computers and Electronics in Agriculture 64(1):39-44.
Sprecher, D. J., D. E. Hostetler & J. B. Kaneene. 1997. A lameness scoring system that uses posture and gait to predict dairy cattle reproductive performance. Theriogenology 47(6):1179-1187.
Thomsen, P. T. 2009. Rapid screening method for lameness in dairy cows. Veterinary Record 164(22):689-690.
Van Hertem, T., V. Alchanatis, A. Antler, E. Maltz, A. Schlageter Tello, et al. 2013. Comparison of segmentation algorithms for cow contour extraction from natural barn background in side view images. Computers and Electronics in Agriculture 91(2):65-74.
Van Hertem, T., E. Maltz, A. Antler, A. Schlageter Tello, C. Lokhorst, et al. 2013. Lameness detection based on multivariate continuous sensing of milk yield, rumination and neck activity. Journal of Dairy Science 96(7):4286-4298.
van Straten, M., I. Siani & D. Bar. 2011. Reduced test-day milk fat percentage in cows diagnosed with claw horn lesions during routine claw trimming. Journal of Dairy Science 94(4):1858-1863.
Viazzi, S., C. Bahr, A. Schlageter Tello, T. Van Hertem, C. E. B. Romanini, et al. 2013. Analysis of individual classification of lameness using automatic back posture measurement in dairy cattle. Journal of Dairy Science 96(1):257-266.
Viazzi, S., C. Bahr, T. Van Hertem, A. Schlageter Tello, C. E. B. Romanini, et al. 2014. Comparison of a three-dimensional and two-dimensional camera system for automated measurement of back posture in dairy cows. Computers and Electronics in Agriculture 100(1):139-147.
Winckler, C. & S. Willen. 2001. The reliability and repeatability of a lameness scoring system for use as an indicator of welfare in dairy cattle. Acta Agriculturae Scandinavica Section A. Animal Science(Supplementum 30):103-107.

List of Tables

Table 1: Classification of n = 744 independent cow-observations represented with confusion matrices. The strict correct classification rate for each confusion matrix is presented below each table.

	(a) Maximising correct classifications					**(b)** Minimising misclassifications					**(c)** Rescaling				
	T1*	T2*	T3*	T4*		T1*	T2*	T3*	T4*		T1*	T2*	T3*	T4*	
	0.16	0.21	0.25	0.30		0.14	0.24	0.26	0.30		0.17	0.21	0.25	0.29	
	Reference					**Reference**					**Reference**				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Model **1**	190	88	15	4	1	49	14	2	1	0	231	133	26	6	1
2	88	154	60	26	4	241	249	104	36	13	47	109	50	24	4
3	12	21	34	9	9	0	1	7	6	1	12	22	35	9	9
4	1	1	8	11	3	1	0	4	7	3	1	0	5	10	2
5	0	0	0	0	5	0	0	0	0	5	0	0	1	1	6
Correct Classification Rate				**53.0**	%				42.6	%				52.6	%

*T1, T2, T3 and T4 represent the cut-off threshold values for transforming the continuous algorithm output to a discrete 5-point scale in each categorisation.

Table 2: The 5-point confusion matrix of the bootstrap voted ordinal multinomial logistic regression model with four consecutive measurements of n = 186 individual cows.

		Reference				
		1	**2**	**3**	**4**	**5**
Model	**1**	47	18	4	1	0
	2	14	48	16	5	1
	3	2	5	10	3	3
	4	0	1	1	4	0
	5	0	0	0	0	3

Table 3: The strict binary confusion matrix of the bootstrap voted ordinal multinomial logistic regression model with four consecutive measurements of $n = 186$ individual cows.

		Reference	
		Lame	**NotLame**
Model	**Lame**	24	8
	NotLame	27	127

Table 4: Classification output of three regression models with consecutive measurements. In the analysis, four consecutive measurements of n = 186 individual cows were used.

	Classification output							
			5-point				Binary	
	(a) Strict	(b) Tolerant	(c)	(d)	(e)		(f)	
	CCR	CCR	MAE	RMSD	CC	CCR	Sensitivity	Specificity
Ordinal Multinomial Logistic Regression	60.2	90.9	0.500	0.852	0.791	81.2	47.1	94.1
Nominal Multinomial Logistic Regression	60.2	91.4	0.511	0.842	0.800	80.7	47.1	93.3
Linear Regression	56.5	91.9	0.522	0.839	0.694	80.7	54.9	90.4

Strict = exact agreement between reference and model;
Tolerant = acceptable one single unit difference between model and reference;

Measures of performance:

MAE = mean absolute error [score unit];
RMSD = root mean squared deviation [dimensionless];
CC = contingency coefficient [dimensionless];
CCR = correct classification rate or model accuracy, the overall ability to correctly classify Lame and Not-Lame animals [%];
Sensitivity = the ability to detect Lame animals [%];

Specificity = the ability to detect Not-Lame animals [%];

List of figures

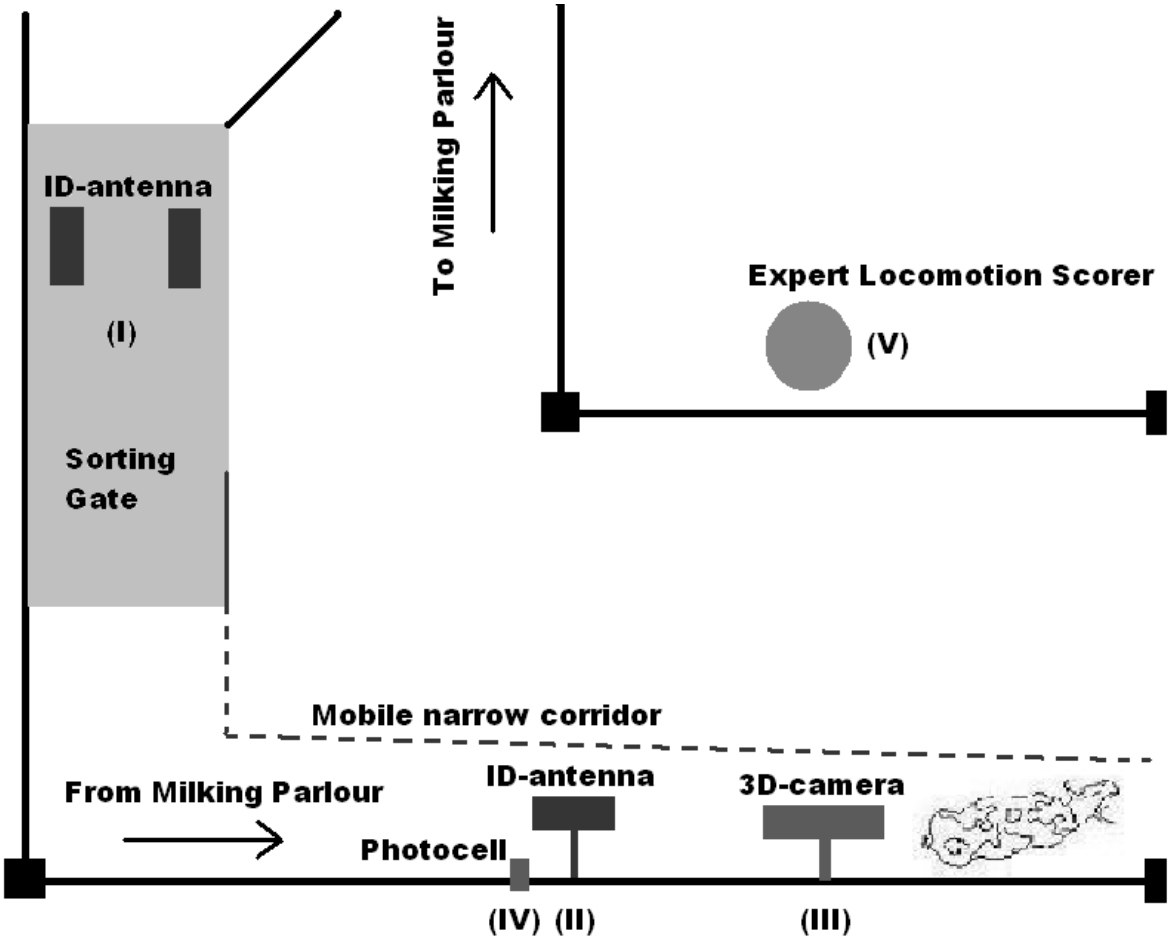

Figure 1: Top view layout of the 3D-camera setup. Data were collected when cows returned from the milking parlour to their cowsheds. All cows were electronically identified by the antennas in the sorting gate (I) and in the corridor (II). The 3D-camera (III) was triggered by the photocell (IV) that was located in the corridor. A live scoring expert (V) performed a locomotion scoring of the cows while they passed through the setup.